SkyEyeGPT: Unifying Remote Sensing Vision-Language Tasks via Instruction Tuning with Large Language Model

Year: 2024 | *Citations: 102* | *Authors: Yangfan Zhan, Zhitong Xiong, Yuan Yuan*

Abstract

Large language models (LLMs) have recently been extended to the vision-language realm, obtaining impressive general multi-modal capabilities. However, the exploration of multi-modal large language models (MLLMs) for remote sensing (RS) data is still in its infancy, and the performance is not satisfactory. In this work, we introduce SkyEyeGPT, a unified multi-modal large language model specifically designed for RS vision-language understanding. To this end, we meticulously curate an RS multi-modal instruction tuning dataset, including single-task and multi-task conversation instructions. After manual verification, we obtain a high-quality RS instruction-following dataset with 968k samples. Our research demonstrates that with a simple yet effective design, SkyEyeGPT works surprisingly well on considerably different tasks without the need for extra encoding modules. Specifically, after projecting RS visual features to the language domain via an alignment layer, they are fed jointly with task-specific instructions into an LLM-based RS decoder to predict answers for RS open-ended tasks. In addition, we design a two-stage tuning method to enhance instruction-following and multi-turn dialogue ability at different granularities. Experiments on 8 datasets for RS vision-language tasks demonstrate SkyEyeGPT's superiority in image-level and region-level tasks, such as captioning and visual grounding. In particular, SkyEyeGPT exhibits encouraging results compared to GPT-4V in some qualitative tests. The online demo, code, and dataset will be released in https://github.com/ZhanYang-nwpu/SkyEyeGPT.